frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 18 November 2010
doi: 10.3389/fpsyg.2010.00174

Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords

Emmanuel Keuleers*, Kevin Diependaele and Marc Brysbaert

Department of Experimental Psychology, Ghent University, Ghent, Belgium

Edited by:
Jonathan Grainger, Centre National de la Recherche Scientifique, France

Reviewed by:
Arnaud Rey, Centre National de la Recherche Scientifique and Aix-Marseille Université, France
Ludovic Ferrand, Centre National de la Recherche Scientifique and Université Blaise Pascal, France
Melvin Yap, National University of Singapore, Singapore

****Correspondence:***
Emmanuel Keuleers, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium.
e-mail: emmanuel.keuleers@ugent.be

In recent years, psycholinguistics has seen a remarkable growth of research based on the analysis of data from large-scale studies of word recognition, in particular lexical decision and word naming. We present the data of the Dutch Lexicon Project (DLP) in which a group of 39 participants made lexical decisions to 14,000 words and the same number of nonwords. To examine whether the extensive practice precludes comparison with the traditional short experiments, we look at the differences between the first and the last session, compare the results with the English Lexicon Project (ELP) and the French Lexicon Project (FLP), and examine to what extent established findings in Dutch psycholinguistics can be replicated in virtual experiments. Our results show that when good nonwords are used, practice effects are minimal in lexical decision experiments and do not invalidate the behavioral data. For instance, the word frequency curve is the same in DLP as in ELP and FLP. Also, the Dutch–English cognate effect is the same in DLP as in a previously published factorial experiment. This means that large-scale word recognition studies can make use of psychophysical and psychometrical approaches. In addition, our data represent an important collection of very long series of individual reaction times that may be of interest to researchers in other areas.

Keywords: lexical decision, visual word recognition, reaction time, practice effect, megastudy, Dutch, word frequency, pseudowords

LARGE-SCALE DATABASES OF WORD PROCESSING TIMES

In recent years, psycholinguistics has seen a remarkable growth of research based on the analysis of data from large-scale studies of word recognition, in particular lexical decision and word naming. Because such databases comprise a substantial part of the lexicon, they can be used to test broad hypotheses about language processing, certainly when they are combined with linguistic resources, such as the CELEX lexical database (Baayen et al., 1995) or other recently developed frequency measures (e.g., Brysbaert and New, 2009; Keuleers et al., 2010). As Baayen (2005) writes: "when combined, the linguistic and psychological resources become a particularly rich gold mine for the study of the lexicon and lexical processing". In addition, the availability of behavioral data for large numbers of words allows researchers to quickly evaluate new hypotheses by simply analyzing the dataset.

Unfortunately, the number of large datasets currently available is very limited, because collecting behavioral data involves a substantial investment. As a matter of fact, besides a few smaller-scale datasets (discussed below), only two very large databases have been published so far. The first is the *English Lexicon Project* (ELP; Balota et al., 2007), which involved the naming of 40,000 words and lexical decisions to the same words. The second study is the *French Lexicon Project* (FLP; Ferrand et al., 2010), which involved a lexical decision task on more than 38,000 words and the same number of nonwords.

In this paper, we present a third large-scale lexical decision study on more than 14,000 Dutch mono- and disyllabic words and an equal number of nonwords. This study differs from the two studies mentioned above, because it involved a sample of 39 participants responding to all stimuli (in different sessions). This considerably improves the analyses that can be done on the data, but raises the question to what extent behavioral word recognition data of the first few sessions are similar to those of later sessions. If there are no big differences, the approach offers an interesting alternative for future studies (also because it is logistically easier to organize). Conversely, if lexical decision performance differs dramatically after a few hours of training, it becomes interesting to investigate what causes these differences and what implications they have for the current practice of building word theories entirely on 1-h experiments.

Before we present the new study, we sketch how large-scale word recognition studies gained prominence in psycholinguistic research. Following Seidenberg and Waters (1989), these studies are often called "megastudies." They involve the presentation of a large sample of unselected stimuli, instead of the small samples of carefully selected stimuli used in the more familiar factorial designs.

A REVIEW OF THE MEGASTUDY APPROACH

As far as we were able to trace, the first megastudy was run by Seidenberg and Waters (1989). A group of 30 students named 2,897 monosyllabic words. Interestingly, this study was never published in a scientific journal[1], but the data are well-known because several groups of researchers have analyzed and reanalyzed them. One of these analyses was published by Treiman et al. (1995). These authors examined whether the rime of English

[1] Seidenberg made them available on his website.

monosyllabic words has an effect on word naming times beyond that of the constituting phonemes, based on the observation that in English there is more regularity in the pronunciation of the complete rime than in the pronunciation of the vowels. Treiman et al. (1995) not only made use of the "more traditional small-scale [factorial] experiments" (p. 108), but also ran a study in which 27 participants named 1,327 consonant–vowel–consonant monosyllabic words. The results of this large-scale study were compared to those of Seidenberg and Waters (1989). In both datasets Treiman et al. (1995) found the predicted effect of rime consistency and further observed that 38–49% of the variance in naming times could be explained by letter-sound consistency, word frequency, word length, neighborhood size (i.e., the number of words of the same length that differed by one letter only), and the nature of the initial phoneme. In particular, the contribution of the last variable was unexpectedly large [5% unique variability in the Treiman et al. (1995) study and 23% in the Seidenberg and Waters (1989) study, which was more than any of the other variables].

Spieler and Balota (1997) shortly afterward published a third large-scale study of word naming, to test how well the times were in agreement with the processing times predicted by two well-known computational models of word naming (Seidenberg and McClelland, 1989; Plaut et al., 1996). Spieler and Balota (1997) argued that because computational models provided processing times for individual words, it was interesting to see how well these predictions correlated with actual word naming latencies. They collected naming times from 31 participants for the 2,820 single-syllable words that had been used in the training corpora of the models. Surprisingly, the naming times correlated less well with the models' predictions (R^2s of 0.03 and 0.10) than with a simple combination of word frequency, word length, and orthographic neighborhood size ($R^2 = 0.22$). In line with Treiman et al. (1995), Spieler and Balota (1997) also noticed that 20% of unique variance in naming latencies was explained by features of the initial phoneme. Importantly, regression analyses of large-scale naming studies started to play a vital role in psycholinguistic discussions (e.g., Balota and Spieler, 1998; Seidenberg and Plaut, 1998; Kessler et al., 2002).

The megastudy approach got further impetus when Balota et al. (2004) added lexical decision times to the naming times. They asked 30 young adults and 30 older adults to make lexical decisions to 2,906 monosyllabic words and 2,906 length-matched pronounceable nonwords (constructed by changing one to three letters in a corresponding word). Using these data, Balota et al. (2004) found that the first phoneme was next to unimportant in lexical decision, but that there were strong effects of word frequency (both objective and subjective), letter-sound consistency, neighborhood size, and semantic variables. Together, these variables accounted for 49% of the variance in the lexical decision times in young adults, and 39% of the lexical decision times in older adults (Balota et al., 2004, **Table 5**; see also Cortese and Khanna, 2007). The accuracy scores were also well predicted by the same set of variables ($R^2 = 0.31$ and 0.20 for young and old adults). This established the usefulness of the megastudy approach.

Encouraged by the above findings, Balota and colleagues embarked on an even more ambitious project: the collection of naming times and lexical decision times for over 40,000 English words (and nonwords), which were no longer limited to single-syllable stimuli. This required a total of 444 participants for the naming experiment (yielding on average 24 observations per word) and a group of 816 participants for the lexical decision experiment (on average 27 observations per word and nonword). The enterprise became known as the *English Lexicon Project* (ELP; Balota et al., 2007; available at http://elexicon.wustl.edu). The ELP database for the first time allowed researchers to do regression analyses on words beyond the single syllable (Kello, 2006; Yap and Balota, 2009; Perry et al., 2010). Around the same time, Lemhöfer et al. (2008) expanded the megastudy approach to the domain of bilingualism, when they administered a progressive demasking study on 1,025 monosyllabic English words to native speakers of French, German, and Dutch.

With the transition from a naming study of 2,000 words to a lexical decision study of more than 40,000 words and the same number of nonwords, Balota et al. (2007) made a methodological change. Whereas in the previous megastudies every participant responded to all stimuli, different groups of participants now responded to different stimuli. Balota et al. (2007) estimated that participants could produce stable data for approximately 2,500 words in the naming task, and 3,500 stimuli in the lexical decision task, distributed over two experimental sessions with not more than a week in-between. As indicated above, this required hundreds of participants. The same approach was followed in the FLP (Ferrand et al., 2010), where participants took part in two sessions of 1 h (for a total of 2,000 observations per participant). This study required 975 participants to get an average of 22 observations per word.

THE ADVANTAGES OF COMPLETE DESIGNS

Although ELP and FLP have proven their worth, they also have their limitations. One of the strongest is that there is no orthogonal variation of participants and stimuli: Each stimulus was processed by a different, randomly selected, group of participants. This introduces noise and complicates statistical analyses if one wants to generalize across participants. For instance, it is not possible to run the traditional F1 analysis of variance with participants as random variable. It also makes the estimation of the participants' effect less precise in the more recent mixed effects analyses. For these reasons, authors who are interested in monosyllabic words tend to prefer the Balota et al. (2004) datasets over the ELP dataset (e.g., Cortese and Khanna, 2007).

A further limitation of ELP and FLP concerns the logistics involved in running an experiment that limits the number of stimuli to 2,000–3,500 per participant. The ELP study was a combined effort of six universities. Similarly, for the FLP-study a total of 975 participants had to be found who were willing to take part in two separate sessions (to give but one example of the costs involved, 62 participants failed to show up for the second session and had to be replaced).

Because of the above limitations, the question arises whether it is possible to run a word recognition study in which a limited group of participants is tested for a prolonged period of time, as is often done in psychophysical and psychometrical research. Are there theoretical reasons why such an approach would be prohibited? Can participants no longer return valid lexical decision times after a few hours of training, or is the limitation to 2-h experiments simply the result of practical considerations?

We could find only two studies that specifically addressed training effects in lexical decision tasks, both with rather encouraging results. Grainger and O'Regan (1992) ran three masked-priming experiments on bilingual language processing, in which they were the only participants. In total, 30 words and 30 nonwords were presented 90 times during 30 sessions of 15 min each, which the authors completed at a rate of two sessions per day. Interestingly, there was no evidence that the results, because of practice effects, at the end of the study differed from those at the beginning, leading the authors to conclude (p. 334) that: "Psychophysical methodology, usually reserved for the study of low-level perceptual processes, therefore appears to hold some promise for investigators of higher-level cognitive phenomena." A similar approach was taken by Ziegler et al. (2000), who worked with eight "well-trained participants" in all their experiments.

The second study (Murray and Forster, 2004, Experiment 3) also involved the two authors as participants, together with a research associate. Each participant was shown the same list of stimuli in three sessions 1 week apart. The frequency effect investigated remained the same throughout the experiment, although there was a training effect on the overall response speed (the three trained participants were faster than the untrained participants of Experiment 1).

So, there would seem to be no *a priori* reasons why participants are not capable of providing valid lexical decision times beyond the first few hours, at least if relevant task considerations are taken into account. The most important consideration for lexical decision arguably is that participants must not be able to detect systematic differences between the words and the nonwords other than the fact that the former are part of the language and the latter not. Otherwise, the participants could pick up the cues to help them make word/nonword decisions, even when they are not aware of these cues (similar to what is observed in implicit learning; Reber, 1989). The more practice participants have with the stimuli of an experiment, the more likely they are to be influenced by unintended differences between the words and the nonwords. An example of such an unintended cue was reported by Chumbley and Balota (1984). In their second experiment, the nonwords were on average one letter shorter than the words, and only weak effects of the semantic predictor variables were found. When the bias was corrected for in Experiment 3, much clearer effects emerged.

THE DUTCH LEXICON PROJECT

On the basis of the above considerations, we decided to run a study in which participants had to respond to over 14,000 mono- and disyllabic words and over 14,000 nonwords. We limited our stimuli to these words, because they comprise nearly all the stimuli used in word recognition research over the last 50 years. We projected that the entire study would take 16–17 h per participant. Below we describe how the study was implemented and we examine whether the outcome was successful. To facilitate reference to the study and to stress its connection to ELP and FLP, we called it the *Dutch Lexicon Project* (DLP).

MATERIALS AND METHODS

PARTICIPANTS

A total of 39 participants finished the experiment. Four more started but were excluded after a few hours. Participants were fully informed about the length of the experiment and about the fact that they would be excluded if their accuracy dropped below 85% in three successive blocks. Performance was monitored at the end of each day. Participants were informed that there were 57 blocks of 500 trials and that, if they completed the experiment successfully, they would receive 200 Euro for the entire experiment. They were also informed that if they dropped out of the experiment or if their accuracy fell consistently below the 85% benchmark, they would only be paid 5€ per hour completed. This reward scheme was set (a) to motivate participants to continue up to the very end of the experiment, and (b) to discourage participants from drawing out the experiment with slow responses (which could be an issue if participants were paid on an hourly basis). Participants were 7 male and 32 female students and employees from Ghent University, ranging in age from 19 to 46 ($M = 23$). After the initial intake with the practice session (see below), participants were free to enter the lab during office hours and to go through the experiment at their own pace, using a booking system to reserve time slots. After reservation, participants could sit at any of the four computers specifically devoted to the study. Upon entering their participation code in the experiment system, they would be presented with their next block of trials. After each completed block, participants could choose whether to continue or to stop the session. The only advice given to the participants was to limit their participation to 2 h per half day. The fastest participant finished the experiment in 10 days time; the slowest took 90 days.

STIMULI

The stimuli comprised 14,089 Dutch words and 14,089 nonwords. Of the word stimuli, 2,807 were monosyllabic, and 11,212 were disyllabic. The base set of words consisted of all mono- and disyllabic word forms with a frequency of 1 per million or higher from the CELEX lexical database (Baayen et al., 1995), excluding words with a space or dash, proper names, one-letter words, and non-infinitive forms of phrasal verbs. To make the database more valuable for research on morphology, we also included mono- and disyllabic word forms with a frequency of less than 1 per million that were inflectionally related to nominal or verbal word forms already in the database. In addition, mono- and disyllabic words with a frequency below 1 in CELEX were included in our study if they appeared in the age-of-acquisition (AoA) norming studies of De Moor et al. (2000) and Ghyselinck et al. (2000), the word association study of De Deyne and Storms (2008) or a list of Dutch–English cognates kindly supplied by Ton Dijkstra.

In the present study, all words were presented exactly once, with one exception: To further investigate the effects of long-term practice, block 1 and block 50 contained the same 500 stimuli in exactly the same order. We decided to repeat Block 1 toward the end of the study to have a more detailed picture of the changes taking place from the beginning to the end of the experiment.

Interestingly, none of the participants seemed to notice the repetition, and the one participant who knew the block was a repetition of a previous block did not have the feeling of having seen the words before.

The nonword stimuli for our experiment were constructed using the Wuggy algorithm (Keuleers and Brysbaert, 2010; available from http://crr.ugent.be/Wuggy). This algorithm generates nonwords by replacing subsyllabic elements of words (onset, nucleus, or coda) by equivalent elements from other words. For instance, given the information that the words *house* and *couch* exist in English, and that dividing them into subsyllabic elements would give *h-ou-se* and *c-ou-ch*, the algorithm is able to create the pseudowords *h-ou-ch* and *c-ou-se*. The algorithm only generates nonwords based on words with the same number of syllables. So, the disyllabic nonwords are based on subsyllabic elements occurring in disyllabic words only. In addition, we tried to pick nonwords that matched the words as closely as possible on a number of criteria. First, each nonword stimulus was matched to its word on its subsyllabic structure. So, the word *br-oo-d* (bread) gave rise to the nonword *sp-oo-d* (two letters in the onset, two letters in the nucleus, one letter in the coda), whereas the word *b-oo-rd* (edge) resulted in the nonword *l-oo-rd* (one letter in the onset, two letters in the nucleus, two letters in the coda). This restriction additionally guaranteed that the nonwords equaled the words on length. Second, we matched the primary stress pattern (initial or final) of the disyllabic words, meaning that the generator only used words with a matching stress pattern as the basis for disyllabic nonwords. Third, we matched each word with a nonword differing on one subsyllabic element per syllable. This criterion removed the possible confound between word length and word likeliness (disyllabic words with only a single change would have resulted in nonwords that resembled the original word more than monosyllabic words with a single change). This manipulation worked for all but 88 of the monosyllabic words, in which two subsyllabic elements had to be changed instead of one. Of the disyllabic nonwords, 19 (0.002%) were constructed by changing three instead of two elements. Fourth, we changed the subsyllabic element that resulted in the smallest deviation in transition probability (calculated separately for the mono- and the disyllabic word forms of the CELEX lexical database). For instance, the word *gr-oe-n* (green) was changed into the nonword *kr-oe-n*, because the replacement of the onset *gr-* by *kr-* resulted in the smallest possible deviation in transition probability from the original word. This manipulation made sure that the transition frequencies could not be used as a cue for word/nonword discrimination. Fifth, we aimed at minimizing the difference in the number of word neighbors between the words and the nonwords. Since we could not optimize all criteria simultaneously, and because one nonword could be the ideal match for several words, we let the program generate five close-matching nonwords, from which the first author picked one. In addition, the handpicking was used to select nonwords that retained the morphological structure of the word as much as possible. For instance, if the word was inflected or derived (e.g., *motors* [motors]),a pseudo-inflected or pseudo-derived nonword was preferred as well (e.g., *rotars*).

DESIGN

The experiment started with an intake session, in which participants received information about the particulars of the experiment, completed a questionnaire about their reading behavior and knowledge of languages (most students in Belgium are multilingual), and ran a practice session of 200 trials. This practice session contained 100 three-syllable words and 100 three-syllable nonwords. It allowed us to demonstrate the main features of the experiment.

The experiment consisted of blocks (one practice block and 58 test blocks). The practice block contained stimuli of three syllables, the test blocks stimuli of one and two syllables. The practice block consisted of 200 trials, the test blocks of 500 trials, except for the last one, which only had 178 trials.

A trial consisted of the following sequence of events. First, two vertical fixation lines appeared slightly above and below the center of the screen, with a gap between them wide enough to clearly present a horizontal string of letters. Participants were asked to fixate the gap as soon as the lines appeared. Five hundred milliseconds later the stimulus was presented in the gap with the center between the vertical lines; the vertical lines remained on the screen. The stimulus stayed on the screen until the participant made a response or for a maximum of 2 s. Participants used their dominant hand for word responses and their non-dominant hand for nonword responses (using response buttons of an external response box connected to one of the USB ports). After the response, there was an interstimulus interval of 500 ms before the next trial started. The screen was blank in this interval.

A block of 500 trials took about 15 min to complete. Because pilot testing showed that this was too long to complete in one go, after every 100 trials the presentation was paused and waited for the participant to press on the space bar. This gave the participants information about their progress in the block, and also gave them the opportunity to take a break if needed.

After each block, participants received feedback about the percentage of correct trials. They were told they should try to aim as high as possible and that their participation would end if they consistently scored below 85% correct responses.

As indicated above, after the intake participants were free to organize the running of the experiment themselves. They signed up to one of the computers, entered their registration code, and automatically started with the next block. Programming was done in C, making use of the Tscope library (Stevens et al., 2006).

RESULTS AND DISCUSSION

The analyses presented here are primarily aimed at the question whether the prolonged practice had strong effects on the lexical decision data. We will address this question in three different ways. First, we look directly at the practice effect: Are the effects of important word variables different at the end of the experiment than in the beginning? Second, we compare the DLP data to the ELP and FLP data: Are the data of the databases comparable? Finally, we examine whether it is possible to replicate well-established findings from the literature with our database: Would authors using the DLP data come to same conclusions as the ones based on the original experiments?

Although the DLP-data are available at the individual-trial level, for most purposes it is more interesting to have summary data. So, for each stimulus two dependent variables were defined: (1) the percentage correct responses (PC, calculated on all 39 participants), and (2) the reaction times (RTs) of the correct responses. RTs from trials on which the stimulus was not correctly identified (10% of the data) or on which an RT below 200 ms or above 1500 ms was registered (0.9% of the data) were not included in the computation of RT averages. The data of block 50 (i.e., the repetition of test Block 1) are not included in these summary measures. This is also true for all analyses reported below, except for the ones dealing specifically with this block.

THE EFFECTS OF PRACTICE

Figure 1 displays the effect of practice on RT and PC over the total duration of the experiment (i.e., all 58 test blocks). As indicated above, Block 50 is a replication of Block 1. Two observations are noteworthy. First, although the practice effect is highly significant ($p < 0.01$), in practical terms it is rather small (some 40 ms difference in RT and 2% difference in PC). This may be partly due to the rather long practice session (200 trials) and to the fact that most of our participants had taken part in lexical decision experiments before. The second observation is that the repetition of Block 1 at Block 50 did not result in a massive learning effect: Mean word accuracy for Block 50 was about 5% higher than for Block 1, but only 2% higher than subsequent blocks in the experiment, and although reaction times for both words and nonwords were faster than for Block 1, they were not unusually fast relative to subsequent blocks in the experiment. The average correlation per participant between the reaction times of Block 1 and the reaction times for the same items repeated in Block 50 was 0.22. Corrected for length using the Spearman-Brown Formula, this gives a reliability of (39 × 0.22) (1 + 38 × 0.22) = 0.92. This is higher than the overall reliability of the RTs, calculated on the basis of the RTs of the first 20 participants and the RTs of the last 19 participants, which corrected for length is (2 × 0.65)/(1 + 0.65) = 0.79. The higher correlation for the repeated block can be expected given that the same words were seen in the same order by the same participants, but at the same time it is further proof that participants responded consistently throughout the experiment.

To test for interactions between practice and lexical characteristics, we first computed *z*-scores of RTs per participant and per block to remove effects due to block differences and variability between participants (as recommended by Faust et al., 1999). We then analyzed the *z*-scores using a mixed effects model, fitting the interaction of block with word frequency, word length and OLD20 (fixed effects) simultaneously with random intercepts for words (cf. Pinheiro and Bates, 2000). This was done in R (R Development Core Team, 2009), using the lme4 package (Bates and Maechler, 2009). We used the SUBTLEX word frequencies, because these correlated best with the behavioral data (for more details, see Keuleers et al., 2010). OLD20 is a measure of orthographic similarity and calculates the minimum number of letter changes needed to transform the target word into 20 other words (Yarkoni et al., 2008). To capture non-linear effects of the predictors we fitted a restricted cubic spline function instead of a simple linear function. A cubic spline combines a number of cubic polynomials defined over a corresponding number of predictor intervals with smooth knot transitions. It can be restricted to avoid overfitting for extreme predictor values (cf. Harrell, 2009). The piecewise nature allows for a more realistic fitting of non-linearities than polynomials defined over the full range of predictor values. For each predictor, we limited the number of knots to the smallest number yielding significant effects ($p < 0.05$) for all constituent polynomials. This resulted in six knots for word frequency, four for word length, and three for OLD20 (see also

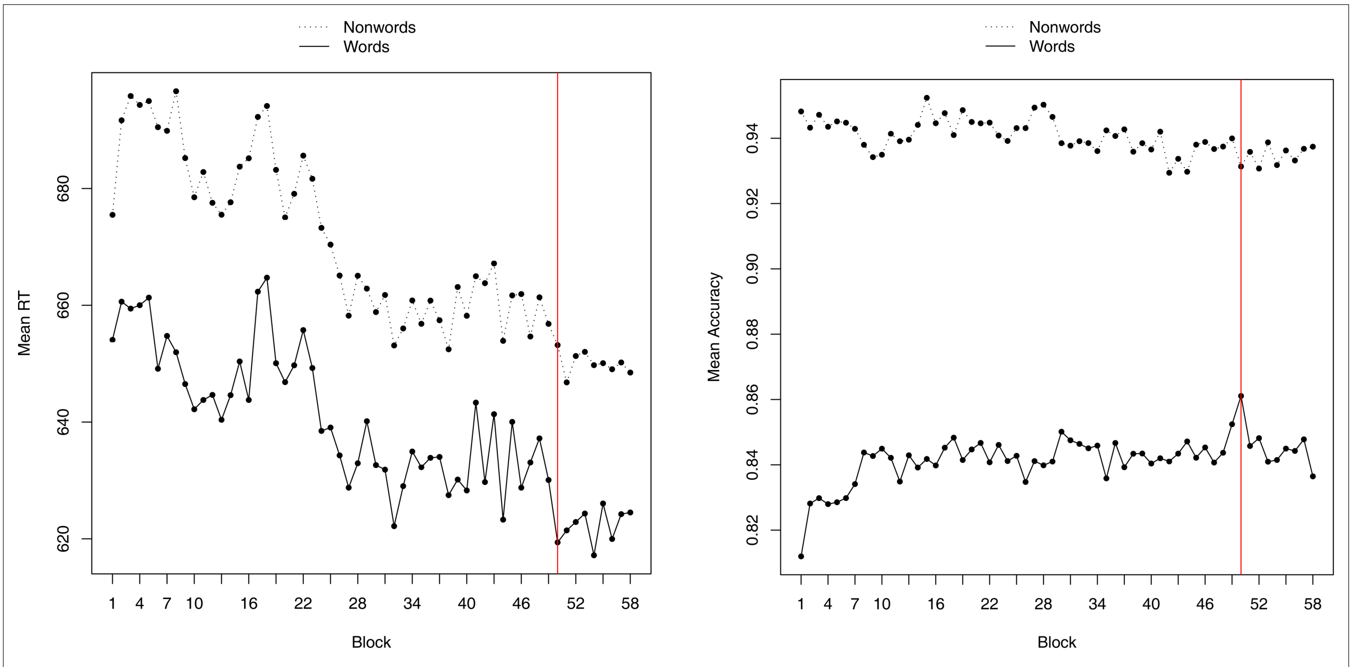

FIGURE 1 | Effects of practice on RTs (A) and PC (B). The full vertical line indicates the 50th block, which was a repetition of Block 1.

Baayen, 2008). For a straightforward evaluation of the practice effects we only considered a linear effect of block. Block 50 was left out of this analysis.

Figure 2 shows the predicted practice effects for the variables frequency, length, and OLD20. Each panel shows the partial effects of the variable at the beginning of the study (after the first block – full line) and at the end of the study (after the last block – dotted line), with the difference between the two lines on the *y*-axis representing the practice effect size in standard deviations (equivalent to Cohen's *d*-measure). The maximum difference in the effect of word frequency and OLD20 is about 0.1 and the maximum difference due to word length is even smaller. Cohen (1992, p. 156) puts the minimum limit of effects sizes at $d = 0.2$, "to be not so small as to be trivial." Still, due to the large number of observations, the mixed effects analysis shows that all interactions with block are significant [block × frequency: $F(5,415351) = 11.93$, MS = 10.06, $p < 0.001$; block × word length: $F(3,415351) = 3.04$, MS = 2.56, $p < 0.03$; block × OLD20: $F(2,415351) = 4.21$; MS = 3.55, $p < 0.02$][2].

To further examine the effect of practice, we performed a separate analysis for each block, this time including block 50, and looked at the linear effects of word frequency, word length and OLD20. Recall that the stimuli in block 1 and block 50, including order, were exactly the same for every individual participant, allowing us to examine whether the repetition across block 1 and block 50 had an effect over and above the effect of task familiarity. For simplicity, we only consider linear effects. **Figure 3** shows that the effects of word frequency and length are clearly stronger in block 1 than in any of the other blocks. This is an important observation, because it suggests that the effect sizes from very short experiments can be inflated, such that some effects may only be found in short experiments. The effect of frequency clearly decreases throughout the experiment, before leveling around block 40. Although the decreasing trend is clear, large frequency effects are found up to the last block. The effect of word length is rather small throughout the experiment, and its evolution is less clear. Word length leads to increased RTs in most blocks, while in some blocks there is an opposite trend. The effect of OLD20 is also rather small throughout the experiment, although it seems to become more important in the second half of the experiment. Block 50 follows the trend for the effect of word frequency quite clearly, but seems to be somewhat removed from the trend for the effects of word length and OLD20, although it is never a clear outlier, indicating that repetition does not seem to have a dramatic influence on effect sizes.

COMPARISON WITH THE ELP AND FLP DATA

A further test of the validity of the DLP dataset is to compare its effects with those of ELP and FLP. **Table 1** lists the main summary statistics of the stimuli together with those of the ELP and FLP. This table clearly illustrates that the words in the DLP were on average shorter and more frequent than those of the other two databases,

[2] *p*-Values are based on MCMC sampling (Baayen, 2008, p. 248).

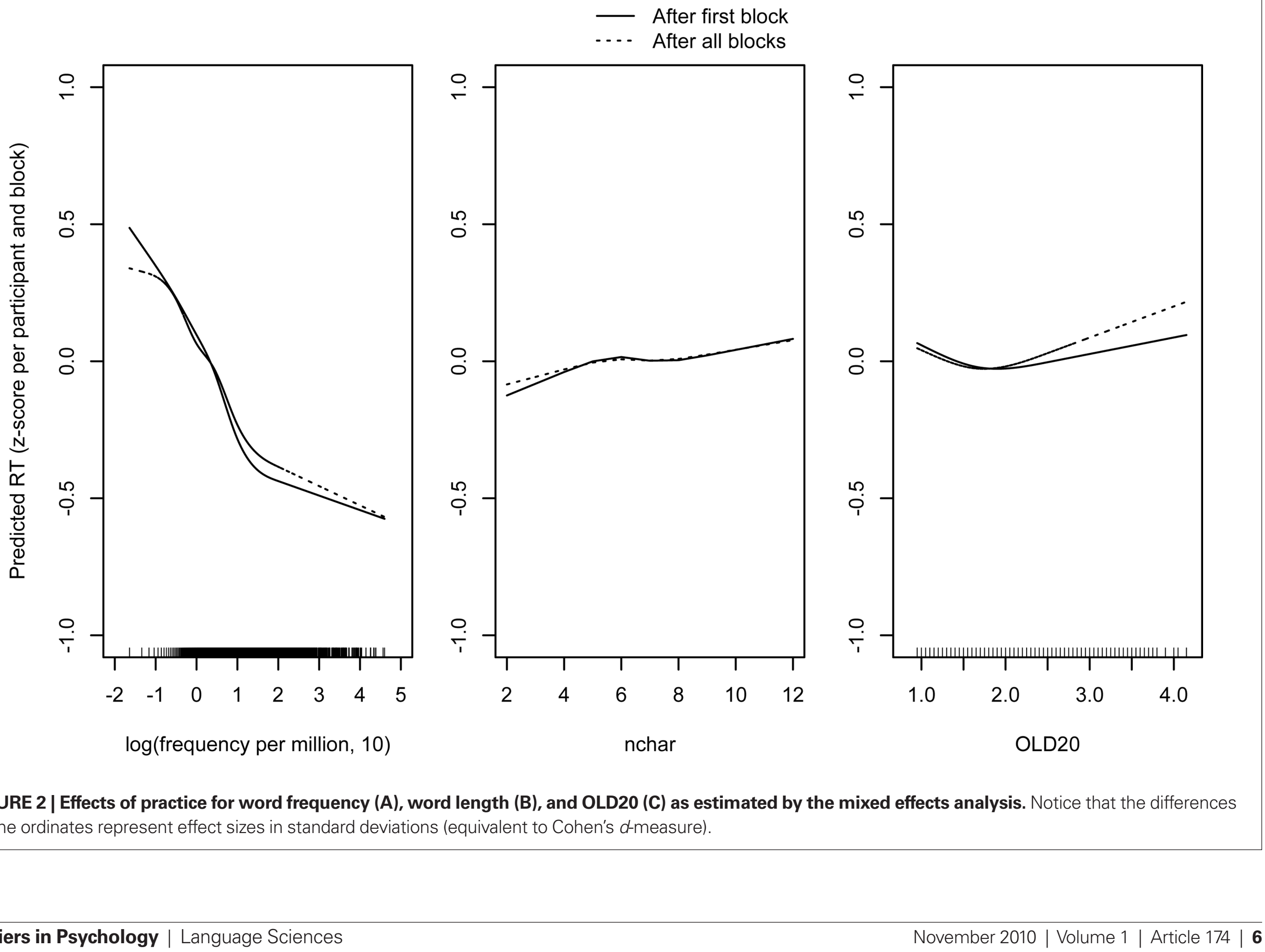

FIGURE 2 | Effects of practice for word frequency (A), word length (B), and OLD20 (C) as estimated by the mixed effects analysis. Notice that the differences on the ordinates represent effect sizes in standard deviations (equivalent to Cohen's *d*-measure).

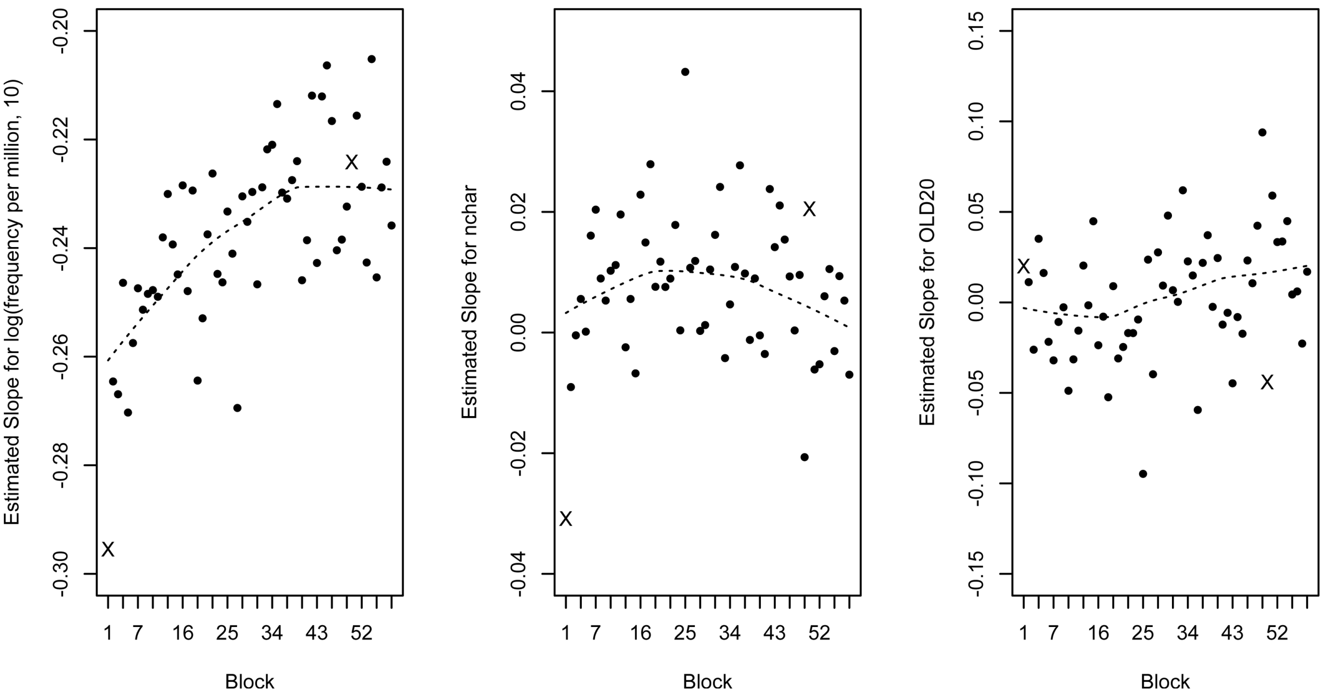

FIGURE 3 | Estimated slopes (in standard deviations) for the linear effects of word frequency (A), word length (B), and OLD20 (C) in each block. The estimates for block 1 and block 50 are marked with an "X." The dotted line shows the locally weighted scatterplot smoothing (LOWESS; Cleveland, 1979), excluding block 50.

Table 1 | Main statistics of the Dutch Lexicon Project, the English Lexicon Project, and the French Lexicon project (between brackets: the range of the variables.)

	DLP	ELP	FLP
Number of words	14,089	40,481	38,840
Length in letters	6.3 (2–12)	8.0 (1–21)	8.5 (2–20)
Length in syllables	1.8 (1–2)	2.5 (1–8)	2.5 (1–7)
Subtitle frequency per million	59.7 (0.02–39,883)	25.2 (0.02–41,857)	21.13 (0–25,988)
Accuracy words	84% (0–100%)	84% (0–100%)	91% (8–100%)
RT words	654 ms (312–1,382 ms)	784 ms (415–1,755 ms)	740 ms (515–1464 ms)
Accuracy nonwords	94% (2–100%)	88% (0–100%)	93% (8–100%)
RT nonwords	674 ms (508–1,135 ms)	856 ms (589–1814 ms)	807 ms (519–1604 ms)

which is to be expected given that DLP was limited to monosyllabic and disyllabic words. **Table 1** also shows that participants were on average 130 ms faster on accepting word stimuli in DLP than in ELP and had the same accuracy level (84%). RTs were 90 ms faster on average in DLP than in FLP, but participants were about 9% less accurate on word decisions. The average frequency in DLP was more than twice that of ELP and FLP. Given that frequency is the most important variable in predicting RTs, this may be the main reason for the longer RTs in ELP and FLP.

To make DLP, ELP, and FLP more comparable, we limited the ELP and FLP data to monosyllabic and disyllabic words, and investigated the impact of well-known predictors on the dependent variables. Among the best predictors of visual lexical decision performance are word length, word frequency, and orthographic similarity to other words. **Table 2** shows the percentages of variance explained by log word frequency, word length (number of characters and number of syllables), and OLD20. Given that the relationships of these variables are not linear (**Figures 5 and 6**; see also New et al., 2006), we checked polynomials up to degree 3.

Table 2 shows that, while the amount of variance accounted for by the different variables is comparable for DLP and FLP, the same variables systematically explain more variance for ELP. In particular the higher correlations with OLD20 and length are striking. Part of the difference is due to the fact that the four predictor variables are interrelated (**Table 3**), so that the stronger frequency effect in ELP has knock-on effects on OLD20 and word length. To correct for this confound, we ran multiple regression analyses. A forward stepwise regression on the DLP RTs with Nchar, Nsyl, Freq, and OLD20 as predictors, returned significant effects of frequency ($\Delta R^2 = 34.4$) and Nchar ($\Delta R^2 = 0.1$). The same analysis for the ELP data returned significant effects for all variables: frequency ($\Delta R^2 = 42.7$), OLD20 ($\Delta R^2 = 3.4$), Nchar ($\Delta\ R^2 = 0.2$), and Nsyl ($\Delta\ R^2 = 0.3$). For the FLP, significant effects were returned for Freq ($\Delta R^2 = 30.7$), OLD20 ($\Delta R^2 = 0.7$), Nsyl ($\Delta R^2 = 0.3$), and Nchar ($\Delta R^2 = 0.1$). These analyses confirm the higher impact of OLD20 in ELP than in DLP and FLP.

A tentative explanation for the higher impact of OLD20 in ELP is the nature of the nonwords. While DLP and FLP had rather sophisticated nonword construction procedures based on recombining subsyllabic elements or trigrams, the ELP nonwords were constructed by changing one or two letters in an existing word. In general, this makes longer stimuli more confusing, because long nonwords often look like a specific word (e.g., captaves-captives), possibly requiring additional processing. In addition, this has the awkward effect that participants often had to answer "yes" to words with very high OLD20 (e.g., "breakthroughs" [OLD20 = 5.75] and "shirtsleeve" [OLD20 = 5.65]) and "no" to nonwords with much lower OLD20s (e.g., "phronological" [OLD20 = 3.75] and "dommunication" [OLD20 = 3.4]).

Table 2 | Percentages of variance accounted for by length in characters (Nchar), length in syllables (Nsyl) frequency (Freq), and orthographic similarity (OLD20) in the Dutch Lexicon Project, the English Lexicon Project (monosyllabic and disyllabic words only; *n* = 22,143), and the French Lexicon Project (monosyllabic and disyllabic words; *n* = 19,184).

	DLP		ELP		FLP	
	RT	PC	RT	PC	RT	PC
Nchar	7.4	1.1	16.4	0.4	6.5	0.3
Nchar poly3	7.5	1.2	16.5	0.5	6.7	0.3
Nsyl	4.7	0.0	9.7	0.4	1.9	0.0
Freq	34.1	18.3	42.7	21.7	30.7	15.1
Freq poly3	35.9	22.4	44.7	25.4	33.7	18.1
OLD20	4.1	0.0	20.0	0.6	5.1	0.6
OLD20 poly3	4.3	0.1	20.3	0.8	5.2	1.0

RT analyses calculated on the zRT scores and accuracy >0.66. Polynomials of the third degree were used to be able to simulate the sigmoid (cubic) frequency curve.

Megastudies not only make it possible to study the impact of a variable on the behavioral data, but also to have a look at the effect across the entire range of the variable. Is there a linear relationship between log frequency and the behavioral data, as authors have assumed for a long time, or are there systematic deviations? And are the effects the same in DLP as in ELP? **Figures 4 and 5** show the effect of word frequency on RTs and PCs in ELP and DLP. To improve the clarity, the word data were grouped in log frequency bins and average values are given, together with error intervals.

Figure 4 shows that the frequency effect on RT is very much the same in DLP, ELP, and FLP, apart from an overall shift in RT. Interestingly, there are considerable differences in the minimal accuracy levels. Whereas the very low-frequency words in DLP had an average accuracy of slightly above 60%, in ELP this was slightly above 70%, and in FLP well above 80%. The most likely origin of this difference again is the nature of the nonwords. In particular, it looks like the nonwords in FLP were easier to detect than in the other two databases.

Figures 4 and 5 further show that the relationship between log frequency and behavioral data is not linear but looks very much like a sigmoid, with a flattening of the curve below 0.1 per million

Table 3 | Intercorrelations between the variables length in characters (Nchar), length in syllables (Nsyl) frequency (Freq), and orthographic similarity (OLD20) for the Dutch Lexicon Project, the English Lexicon Project (monosyllabic and disyllabic words only; *n* = 22,143), and the French Lexicon Project (monosyllabic and disyllabic words; *n* = 19,184).

	DLP			ELP			FLP		
	Nchar	Nsyl	Freq	Nchar	Nsyl	Freq	Nchar	Nsyl	Freq
Nsyl	0.611			0.589			0.455		
Freq	–0.319	–0.279		–0.373	–0.278		–0.343	–0.279	
OLD20	0.676	0.367	–0.254	0.801	0.539	–0.356	0.636	0.412	–0.262

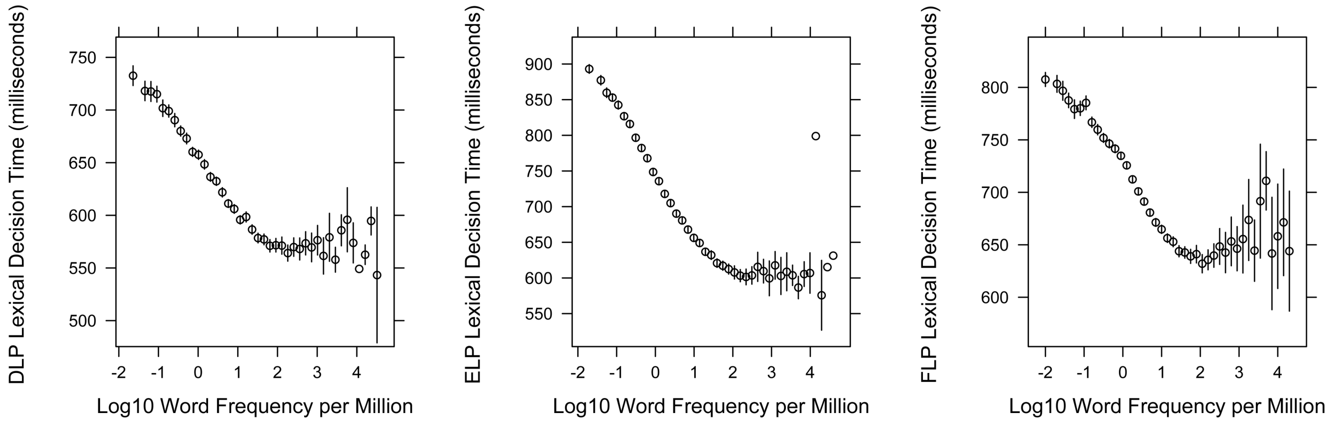

FIGURE 4 | The word frequency-RT curve for the word stimuli in DLP, ELP, and FLP. Stimulus frequencies were obtained from SUBTLEX-NL, SUBTLEX-US, and Lexique 3.55 and varied from 0.02 to nearly 40,000 per million words. Circles indicate the mean RT per bin of 0.15 log word-frequency; error bars indicate 2 × SE (bins without error bars contained only one word).

(log value of −1) and above 50 per million (log value of +1.7). Intriguingly, some very high-frequency words are not responded to well. These are often function words (e.g., articles, prepositions) or parts of fixed expressions (e.g., the French negation "ne pas"; both the words "ne" and "pas" did badly in FLP). A final surprising feature of the curves is that nearly half of the frequency effect is due to frequencies below 1 per million (log value of 0). Ironically, researchers have traditionally tried to increase the frequency effect by selecting very high frequency words (above 100 per million) rather than by selecting very low frequency words (in many studies the low-frequency range is loosely defined as below 5 per million or even 10 per million). On the basis of our data it looks as if the lower half of the frequency curve has been neglected in favor of the less interesting (higher) half.

VIRTUAL EXPERIMENTS

A further test of DLP's usefulness is to examine whether patterns of results found in existing small-scale factorial experiments can be replicated with the dataset. In other words, do the results reported in the original publications hold when we set-up virtual experiments using the same stimuli from the DLP database?[3] This is particularly important when researchers want to use the DLP data to test or verify experimental hypotheses.

One of the classic Dutch visual word recognition studies based on lexical decision was published by Schreuder and Baayen (1997). The authors addressed the question to what extent lexical decision times to singular nouns are influenced by the frequencies of the plurals. For instance, the words *spier* (muscle) and *stier* (bull) have more or less the same frequency, but the plural form *spieren* (muscles) occurs significantly more often than the plural form *stieren* (bulls). Schreuder and Baayen hypothesized that singular nouns with frequent plurals would be responded to faster than matched singular nouns with non-frequent plurals. After confirming this hypothesis, they examined the effect of the number of morphologically related nouns (family size) and the cumulative frequency of all family members (cumulative family

[3] A number of studies could not be simulated because they did not include the stimuli used and the authors were no longer able to find them. This once again underlines the importance of including stimulus materials in articles.

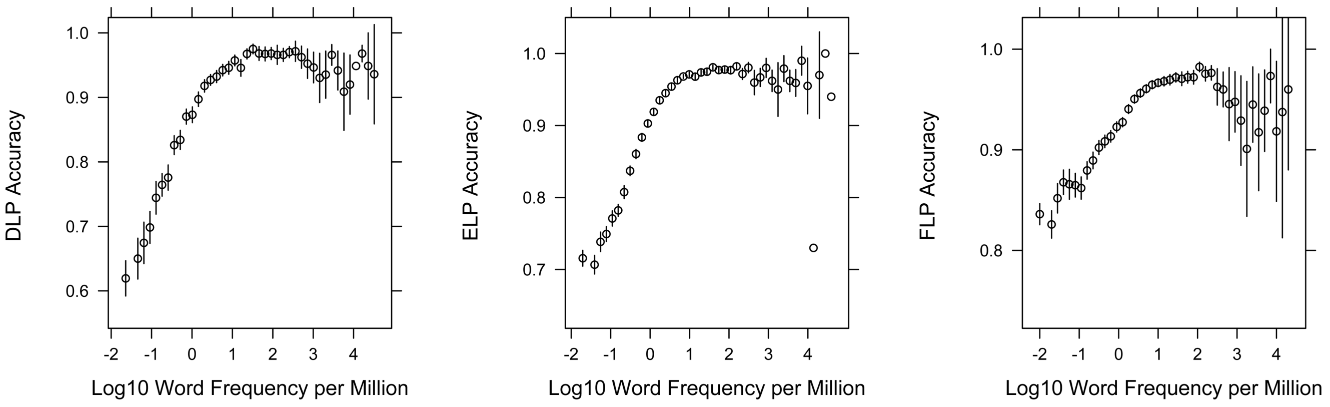

FIGURE 5 | The word frequency-accuracy curve for the word stimuli in DLP, ELP, and FLP. Stimulus frequencies were obtained from SUBTLEX-NL, SUBTLEX-US, and Lexique 3.55, and varied from 0.02 to nearly 40,000 per million words. Circles indicate the mean accuracy per bin of 0.15 log word-frequency; error bars indicate 2 × SE.

Table 4 | Reaction times (in ms) to singular Dutch nouns as a function of the frequencies of the plurals and the family size, as reported by Schreuder and Baayen (1997) and in virtual DLP experiments.

	Original experiments RTs	Virtual experiments RTs	Number of stimuli in virtual experiments (original number of stimuli per condition)
Exp. 1: high vs. low-frequency plurals	539 vs. 580**	583 vs. 631**	35 vs. 34 (35)
Exp. 2: high vs. low cumulative family frequency	599 vs. 644**	606 vs. 665**	27 vs. 27 (32)
Exp. 3: high vs. low family size	553 vs. 594*	599 vs. 646*	18 vs. 16 (18)
Exp. 4: high vs. low cumulative frequency for fixed family size	598 vs. 612	666 vs. 664	16 vs. 11 (17)
Exp. 5: high vs. low singular for fixed family size and cumulative frequency	576 vs. 656**	625 vs. 709**	18 vs. 9 (20)

*For original experiments: **$p < 0.01$, *$p < 0.05$ in both F1 and F2 analysis.*
*For virtual experiments: **$p < 0.01$, *$p < 0.05$ in mixed effects analysis including random intercepts for participants and stimuli.*

frequency). All in all, Schreuder and Baayen ran five experiments, of which the results are summarized in the left column of **Table 4**.

In order to run the virtual experiments, we looked up the RTs of our participants to the stimuli in Schreuder and Baayen's (1997) experiments that were also present in our database. RTs from correct responses between 200 ms and lower than 1500 ms were then analyzed with a mixed effects model, with condition as a fixed effect and crossed random effects of participants and stimuli. **Table 4** shows the results of the virtual experiments. Although the RTs overall were longer than those of Schreuder and Baayen (1997), the pattern of effects was exactly the same. That is, Schreuder and Baayen (1997) would have drawn the same conclusions on the basis of a series of virtual experiments using the DLP database. The main limitation is the absence of some stimuli in the database: Schreuder and Baayen included some trisyllabic words in their experiments and also used some words with a surface frequency that was lower than the range examined in DLP. For Experiment 5 in particular, quite some of the original stimuli were not found in the database, resulting in less than half of the observations included in the second condition (9/20).

Building on Schreuder and Baayen's (1997) work, Baayen et al. (1997) designed an experiment with three variables: (1) the summed frequency of the singular and the plural (plus the diminutive forms), (2) the relative frequencies of the singular and the plural noun, and (3) whether the participants responded to the singular noun or to the plural. The results of this experiment are summarized in the left column of **Table 5**.

Baayen et al. (1997) found significant main effects of all three variables: (1) RTs were faster to words with high summed frequencies than to words with low summed frequencies, (2) RTs were faster to words with higher singular frequencies than to words with lower singular frequencies, and (3) RTs were faster to singular nouns than to plural nouns. Baayen et al. also found a critical interaction between the relative frequencies of the singular and the plural nouns and whether participants responded to a singular or a plural noun. Responses were considerably faster to high frequency plural nouns than to low-frequency plural nouns, whereas the summed frequency determined the RTs to the singular forms. Finally, there was also a significant interaction between the summed frequency and responses to singular vs. plural: The time cost of a plural noun was higher for the words with a low summed frequency than for the words with a high summed frequency.

Table 5 | Reaction times reported by Baayen et al. (1997, Experiment 1) for different types of nouns and the results of the virtual experiments with DLP data.

	Original experiments		Virtual experiments	
	Singular noun	Plural noun	Singular noun	Plural noun
Low summed frequency: freqsing > freqplural	612	708	628	715
Low summed frequency: freqsing < freqplural	606	645	617	640
High summed frequency: freqsing > freqplural	561	615	553	643
High summed frequency: freqsing < freqplural	551	558	562	587

To see whether this pattern would be obtained in a virtual experiment, all data were extracted from the DLP and analyzed with a mixed effects model with crossed random effects for participants and stimuli and fixed effects for the main variables and interactions. Of the 186 items, 182 were found in the database. The main effects that were found to be significant by Baayen et al. (1997) were also significant in the mixed effects analyses of the virtual experiment ($p < 0.01$), as was the crucial interaction between the relative frequencies of the singular and the plural and whether participants responded to a singular or a plural noun ($p < 0.01$). Only the interaction between summed frequency and whether participants responded to a singular or a plural noun was not significant ($p = 0.19$). No other effects were significant. So, in all important aspects a virtual experiment on the basis of DLP would have led to the same conclusions as the data obtained by Baayen et al.

Another important topic in Dutch word recognition research has been to what extent word processing is influenced by knowledge of a second language. The cognate-effect is well known in this respect. Bilinguals have a processing advantage for words with a high form overlap that also have the same meaning in the two languages (e.g., *lamp-lamp* in English and Dutch). van Hell and Dijkstra (2002) reported that Dutch native speakers responded about 30 ms faster to Dutch–English cognates in a lexical decision task than to control words (see the left part of **Table 6**). Interestingly, the effect was much smaller for Dutch–French cognates, arguably because Dutch speakers from the Netherlands have a larger knowledge of English. To test this hypothesis, van Hell and Dijkstra (2002) tested bilinguals with a high proficiency in French (these were students taking a French degree), and found more evidence for a French cognate effect. Surprisingly, for the highly proficient French speakers, the English cognate effect was also larger.

We looked up the stimuli used by van Hell and Dijkstra (2002) in the DLP dataset and analyzed the virtual experiment using a linear mixed effects model. In general, the data agreed quite well

Table 6 | The cognate effect reported by van Hell and Dijkstra (2002). Left part: original data. Right part: Simulations with the DLP data. Between brackets: the number of stimuli found in DLP and the number of stimuli used in the original experiment.

	van Hell and Dijkstra		DLP
	Low French	High French	
Dutch–English cognates	499	489	559 (20/20)
Dutch–French cognates	519	520	585 (17/20)
Control words	529	541	595 (37/40)
English cognate effect	30*	52**	36*
French cognate effect	10	21*	10

*$p < 0.05$, **$p < 0.01$.

with those of the Dutch speakers with low French proficiency (see the right part of **Table 6**). To some extent this is surprising, given that Belgian native speakers of Dutch are much more exposed to French than Dutch native speakers from the Netherlands.

Another important topic of study in Dutch word recognition research is the AoA effect. Brysbaert et al. (2000) published a series of experiments showing that a word frequency effect was still found when words are controlled for length, AoA, and imageability. Similarly, a significant AoA effect was found when all other variables were controlled for. However, no significant effect of imageability was found once the stimuli were controlled for length, frequency, and AoA. **Table 5** shows the findings (left part). As the right part of the table shows, these findings are also obtained in a series of three virtual experiments with the Dutch Lexicon data.

One reason why imageability does not have a significant effect on lexical decision times may be that it is not the right variable. van Hell and de Groot (1998) argued that the concreteness effect reported in lexical decision (which is very closely related to imageability) is an artifact of context availability (CA, i.e., how easily a participant can think of a context in which the word can be used). To investigate the issue, van Hell and de Groot compiled four lists of 20 words. The first two lists compared abstract and concrete words that were matched on CA; the second two compared abstract and concrete words confounded for CA (i.e., the CA was much higher for the concrete than the abstract words). Only in the latter condition did van Hell and de Groot (1998) find a significant difference (see the left part of **Table 8**), making them conclude that the concreteness effect was a CA effect in disguise. As before, the same conclusion was reached on the basis of a virtual experiment (right part).

All in all, it looks very much like the DLP data can be used to replicate classic studies in Dutch visual word recognition and, hence, to test new hypotheses. The main limitation is that ideally we would have included more low-frequency words. As indicated in the method section, we largely limited the words to those that had a base form frequency of 1 per million or more. Given that half of the frequency effect is situated below this value (**Figures 4 and 5**), for future studies it is desirable to include many more low-frequency words, so that a more detailed picture becomes available of what happens at the low end of the frequency curve.

One of the reasons why we were successful in the replication of the above studies, is that most of the effects were of a considerable size (30 ms and more). When we tried to simulate a few small effects (20 ms or less), we usually obtained non-significant trends in the expected direction, certainly when the numbers of stimuli were small. This made us realize that power is an issue in megastudies as much as in small-scale studies. To get an idea of the effects that can be simulated with the DLP database, we performed a Monte Carlo simulation, taking 1000 random samples of two sets of n items, adding d milliseconds to the RTs of the second set of the items. On each sample, we ran a mixed effects analysis with participants and items as random effects, and with fixed effects for log frequency, log frequency squared, and condition (unmodified vs. modified RTs). Obtaining a significant effect for condition in such an analysis means that an effect of size d was found with two samples of n items, controlling for the effect of frequency. We then counted the number of times in which the t value for the effect of condition exceeded 2, a conservative heuristic for obtaining a significance level of 0.05.[4] The simulation was repeated with different set sizes (n = 10, 20, 40, 80, 160, 320, 640, 1280) and effect sizes (d = 0, 5, 10, 20, and 40 ms).

Table 7 | Empirical data reported by Brysbaert et al. (2000) for word frequency, AoA, and imageability. Left part: original data. Right part: Simulations with the DLP data. Between brackets: the number of stimuli found in DLP and the number of stimuli used in the original experiment.

	Brysbaert et al. (2000)	DLP
AoA		
Early	594	584 (23/24)
Late	646	648 (21/24)
Effect	52**	64**
FREQUENCY		
High	554	553 (24/24)
Low	639	642 (17/24)
Effect	85**	89**
IMAGEABILITY		
High	609	598 (23/24)
Low	609	614 (23/24)
Effect	0	16

***$p < 0.01$.*

Table 8 | Results obtained by van Hell and de groot (1998) for abstract and concrete words, when the words were matched on context availability and when they were not. Left part: original data. Right part: Simulations with the DLP data. Between brackets: the number of stimuli found in DLP and the number of stimuli used in the original experiment.

	van Hell and de Groot (1998)		DLP	
	Matched CA	**Confounded CA**	**Matched CA**	**Confounded CA**
Abstract	541	554	564 (19/20)	583 (20/20)
Concrete	554	523	581 (19/20)	540 (17/20)
Difference	–13	31**	–17	43**

Figure 6 illustrates the results of the Monte Carlo simulation. One can reasonably expect to find an effect of 40 ms using about 35 items per set. To find an effect of 20 ms, over 130 items per condition are required. For effects of 10 ms or below, the use of the database becomes impractical unless samples of nearly 500 items per condition are possible.

CONCLUSION

In this article we described the results of a large-scale lexical decision study in Dutch. In line with other large word recognition studies (Balota et al., 2007; Ferrand et al., 2010), this study was not

[4] Mixed effects models do not provide p values (Baayen et al., 2008). Although empirical p values can be obtained using MCMC sampling, that procedure is very time consuming. In this case, the MCMC sampling would need to be performed for forty thousand tests.

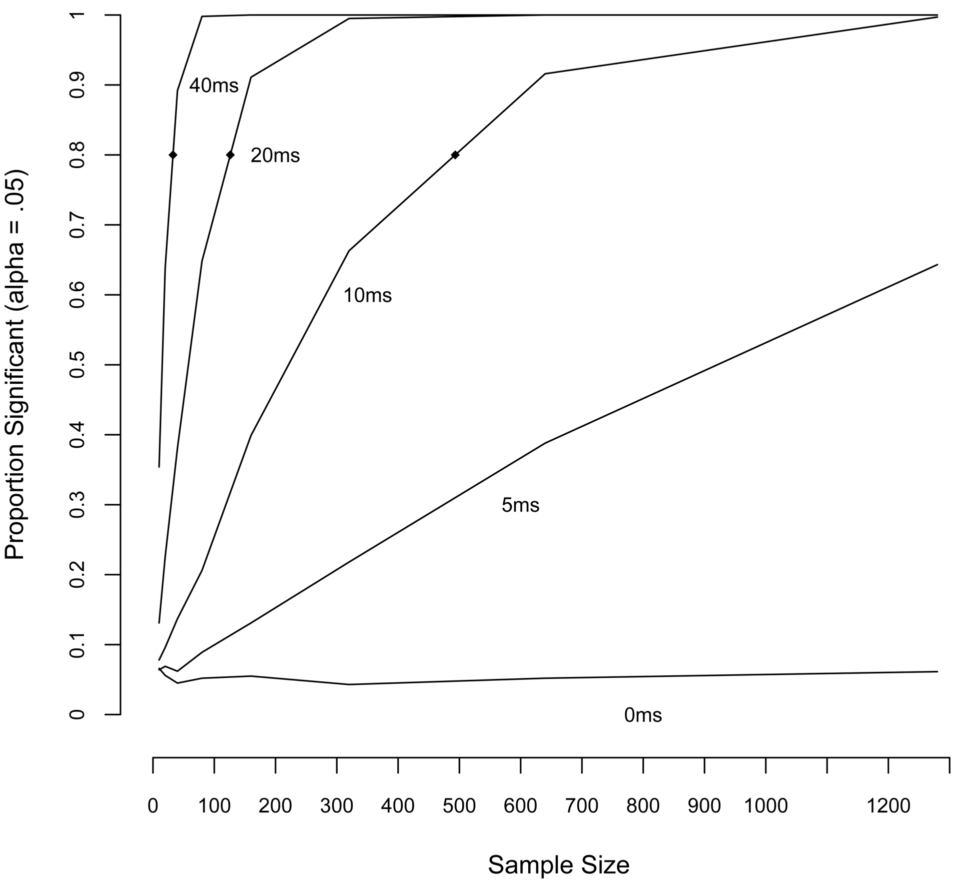

FIGURE 6 | Sample size required for finding an effect of a particular size (in milliseconds), derived by Monte Carlo simulation using two random samples of words from the database. For each combination of sample size (n = 10, 20, 40, 80, 160, 320, 640, 1280) and effect size (0, 5, 10, 20, 40 ms), we ran 1000 simulations. The y-axis indicates the proportion of simulations in which the null hypothesis (no-effect) was rejected. Sample sizes at which sufficient power (0.8) is reached are n = 33 for 40 ms, n = 126 for 20 ms, and n = 493 for 10 ms).

set up with a particular hypothesis in mind, but rather with the aim of making a broad range of word recognition data available, allowing researchers to run regression analyses over the entire range of a variable and to run virtual experiments in order to quickly test a hypothesis.

In contrast with previous large word recognition studies that used many participants responding to a small part of the stimuli, the participants in our study responded to all stimuli. This makes new analyses possible (e.g., about individual differences) and also increases the power of the analyses. For instance, many studies nowadays make use of mixed effects methods to analyze the data. These methods do not rely on average data per stimulus, but take into account participants and items as random effects. Having the same participants for all items allows for less complex interactions between participants and items and, hence, for a better estimation of these random effects. Even for the more traditional statistical methods, having all stimuli responded to by the same participants considerably simplifies matters (e.g., to run an F1 analysis).

Of course, such an approach is only useful if the behavioral data at the end of the experiment are comparable to those at the beginning. Therefore, the analyses in the present paper were focused on this question. We have given three arguments why the approach is fruitful. First, the practice effects are in all respects rather small (**Figures 1–3**). Second, looking at well-established effects, such as the word frequency effect, we see a curve that is very comparable to the one obtained in the other megastudies (**Figures 4 and 5**; see also **Tables 1–3**). Finally, we were able to replicate the core findings of Dutch studies using lexical decisions to printed words (**Tables 4–8**). Our success with virtual experiments contrasts with the disappointing results obtained by Sibley et al. (2009), who failed to replicate the results of several classical experiments concerning the regularity–frequency interaction with the same stimuli from three naming megastudies. Several factors may account for this difference. First, we used linear mixed effects models on trial level data, a method of analysis that may be more powerful than the item-level analysis used by Sibley et al. Second, Sibley et al. looked only at the frequency-regularity interaction in naming, leaving open the possibility that variables related to this particular effect may make it hard to replicate the findings using megastudy data. Third, the authors made no effort to analyze the power of the megastudies to reveal the effects they were interested in. Because the effects are usually greater in lexical decision than in naming, they may be more easy to replicate, a matter that should be investigated further.

We think the main reason why our approach worked, was that we were very careful about the construction of the nonwords. We made every attempt to match the nonwords as much as possible to the words with respect to their sublexical properties, while at the same avoiding too much overlap between the nonwords and the words from which they were derived (which we think was a problem in ELP, because there the nonwords were mostly created by changing a single letter of the word; see Ferrand et al., 2010, for a more extended discussion). Because of these controls, participants had less opportunity to develop implicit learning based on systematic non-lexical differences between the words and the nonwords.

In principle, there are no elements in our data that would prevent us from testing participants on even more trials, which may be necessary given the need for more information on low-frequency words (cf. the missing data in **Tables 4–8**). Whether such a study is feasible in practice, remains, of course, an open question. (How many participants would finish a study of 40 h of lexical decisions?) An alternative may be to break the stimulus set into a small number of equivalent lists and have different groups of participants complete them. This would agree with the traditional split-plot design in analysis of variance and would in all likelihood lead to very similar results provided the participant groups are large enough, so that idiosyncrasies of the participants have relatively little weight.

AVAILABILITY OF THE DLP DATA

The DLP data are available at the trial level and at the item level. Rather than providing a query interface, we are making all data available for direct download in three convenient formats: tab delimited text, R data files, and Excel 2008 spreadsheets, so that researchers have maximal access and flexibility in working with the data. In addition, we are making available a file of stimulus characteristics, which can be merged with the data. All material can be downloaded from http://crr.ugent.be/dlp.

TRIAL LEVEL DATA

At the trial-level, there are 1,098,942 rows of data. For each trial, the following information is given.

- Environment: indicates which of the four computers the participant was using when the trial was recorded.
- Participant: identification number of the participant.
- Block: the number of the block in which the trial was presented.
- Order: the presentation order of the trial for the participant.
- Trial: the trial identification number.
- Spelling: the spelling of the stimulus.
- Lexicality: whether the stimulus was a word (W) or nonword (N).
- Response: the response to the stimulus. Word (W), nonword (N), or time-out (T).
- Accuracy: 1 if the response matched the lexicality, otherwise 0.
- Previous accuracy: accuracy on the previous trial.
- RT: reaction time on the trial.
- Previous RT: reaction time on the previous trial.
- Microsec error: the timing error given by the tscope software (in microseconds).
- Unix seconds: date and time in Unix seconds format (seconds elapsed since 1970).
- Unix microseconds: decimal part of unix seconds (in microseconds).
- Trial day: indicates how many trials the participant responded to since the day began (including the current trial).
- Trial session: indicates how many trials the participant responded to since the session began (including the current trial). A session expired after no response was given for 10 min.
- Order in block: the presentation order of the trial in a block of 500 items.
- Order in subblock: the presentation order of the trial in a subblock of 100 items.

ITEM LEVEL DATA

At the item level, there are 28,178 rows of data. For each stimulus (word or nonword), the following information is given.

- Spelling: the spelling of the stimulus as it was presented.
- Lexicality: whether the stimulus was a word (W) or nonword (N).
- RT: the average reaction to the stimulus.
- Zscore: the average standardized reaction time. Standardized reaction times were calculated separately for all levels of participant, block and lexicality (e.g., all RTs to word-trials in block 1 by participant 1).
- Accuracy: average accuracy for the stimulus.
- RT SD: standard deviation for the average reaction time.
- Zscore SD: standard deviation for the average Zscore.
- Accuracy SD: standard deviation for the average accuracy.

STIMULUS CHARACTERISTICS

- Coltheart N: the number of words of same length differing in one letter, computed over all wordforms in the Dutch CELEX lexical database.
- OLD20: The average Orthographic Levenshtein distance of the 20 most similar words, computed over all wordforms in the Dutch CELEX lexical database.
- CELEX frequency: Raw frequency of the stimulus as given by CELEX.
- CELEX CD: Contextual diversity (dispersion) of the stimulus in CELEX.
- CELEX frequency lemma: sum of the raw frequencies of all possible lemmas for the stimulus in CELEX.
- SUBTLEX frequency: raw frequency of the stimulus in the SUBTLEX-NL database.
- SUBTLEX CD: contextual diversity of the stimulus in SUBTLEX-NL.
- SUBTLEX frequency lemma: sum of the raw frequencies of all possible lemmas for the stimulus in SUBTLEX-NL.
- SUBTLEX frequency million: frequency per million of the stimulus in SUBTLEX-NL.
- SUBTLEX log10 frequency: log10 of raw frequency in SUBTLEX-NL.
- SUBTLEX CD percentage: contextual diversity as a percentage of movies the form occurs in.

- SUBTLEX log10 CD: log10 of contextual diversity in SUBTLEX-NL.
- Summed monogram: sum of non-positional letter frequencies, computed over all wordforms in CELEX.
- Summed bigram: sum of non-positional bigram frequencies.
- Summed trigram: sum of non-positional trigram frequencies.
- Stress: primary stress location (10: initial stress, 01: final stress, 1: monosyllabic).
- Nchar: length of the stimulus in characters.
- Nsyl: length of the stimulus in syllables.
- Morphology: morphological status (e.g., monomorphemic, complex) of the form in CELEX. Different options are separated by a dot.
- Flection: flection (e.g., singular, plural) of the form in CELEX. Different options are separated by a dot.
- Synclass: syntactic class (e.g., Verb, Noun) of the form in CELEX. Different options are separated by a dot.

ACKNOWLEDGMENTS

Part of this research was made possible by an Odysseus grant awarded to Marc Brysbaert by the Government of Flanders.

REFERENCES

Baayen, R. H. (2005). "Data mining at the intersection of psychology and linguistics," in *Twenty-First Century Psycholinguistics: Four cornerstones*, ed. A. Cutler (Mahwah, NJ: Lawrence Erlbaum Associates), 69–83.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412.

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: evidence for a parallel dual-route model. *J. Mem. Lang.* 37, 94–117.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database* (release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *J. Exp. Psychol. Gen.* 133, 283–316.

Balota, D. A., and Spieler, D. H. (1998). The utility of item-level analyses in model evaluation: a reply to Seidenberg and Plaut. *Psychol. Sci.* 9, 238.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English lexicon project. *Behav. Res. Methods* 39, 445–459.

Bates, D., and Maechler, M. (2009). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 0.999375–31. Available at: http://cran.us.r-project.org/web/packages/lme4/index.html

Brysbaert, M., Lange, M., and Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: further evidence from the Dutch language. *Eur. J. Cogn. Psychol.* 12, 65–85.

Brysbaert, M., and New, B. (2009). Moving beyond Ku era and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* 41, 977.

Chumbley, J. I., and Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Mem. Cogn.* 12, 590–606.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74, 829–836.

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159.

Cortese, M. J., and Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: an analysis of 2,342 words. *Q. J. Exp. Psychol.* 60, 1072–1082.

De Deyne, S., and Storms, G. (2008). Word associations: norms for 1,424 Dutch words in a continuous task. *Behav. Res. Methods* 40, 198.

De Moor, W., Ghyselinck, M., and Brysbaert, M. (2000). A validation study of the age-of-acquisition norms collected by Ghyselinck, De Moor and Brysbaert. *Psychol. Belg.* 40, 99–114.

Faust, M. E., Balota, D. A., Spieler, D. H., and Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: implications for group differences in response latency. *Psychol. Bull.* 125, 777–799.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The French Lexicon Project: lexical decision data for 38,840 French words and 38,840 pseudowords. *Behav. Res. Methods* 42, 488.

Ghyselinck, M., De Moor, W., and Brysbaert, M. (2000). Age-of-acquisition ratings for 2816 Dutch four-and five-letter nouns. *Psychol. Belg.* 40, 77–98.

Grainger, J., and O'Regan, J. K. (1992). A psychophysical investigation of language priming effects in two English-French bilinguals. *Eur. J. Cogn. Psychol.* 4, 323–339.

Harrell, F. E. (2009). *Design: Design Package*. R package version 2.3-0.

Kello, C. T. (2006). Considering the junction model of lexical processing. *From Inkmarks to Ideas: Current Issues in Lexical Processing*, ed S. Andrews (New York: Psychology Press), 50–75.

Kessler, B., Treiman, R., and Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *J. Mem. Lang.* 47, 145–171.

Keuleers, E., and Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. *Behav. Res. Methods* 42, 627–633.

Keuleers, E., Brysbaert, M., and New, B. (2010). SUBTLEX-NL: a new frequency measure for Dutch words based on film subtitles. *Behav. Res. Methods* 42, 643–650.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., and Zwitserlood, P. (2008). Native language influences on word recognition in a second language: a megastudy. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 12–31.

Murray, W. S., and Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychol. Rev.* 111, 721–756.

New, B., Ferrand, L., Pallier, C., and Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: new evidence from the English Lexicon Project. *Psychon. Bull. Rev.* 13, 45.

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cogn. Psychol.* 61, 106–151.

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115.

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Available at: http://www.R-project.org

Reber, A. S. (1989). Implicit learning and tacit knowledge. *J. Exp. Psychol. Gen.* 118, 219–235.

Schreuder, R., and Baayen, R. H. (1997). How complex simplex words can be. *J. Mem. Lang.* 37, 118–139.

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568.

Seidenberg, M. S., and Plaut, D. C. (1998). Evaluating word-reading models at the item level: matching the grain of theory and data. *Psychol. Sci.* 9, 234.

Seidenberg, M. S., and Waters, G. S. (1989). Word recognition and naming: a mega study. *Bull. Psychon. Soc.* 27, 489.

Sibley, D. E., Kello, C. T., and Seidenberg, M. S. (2009). "Error, error everywhere: a look at megastudies of word reading," in *Proceedings of the 2009 Meeting of the Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Austin, TX: Cognitive Science Society), 1036–1041.

Spieler, D. H., and Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychol. Sci.* 8, 411–416.

Stevens, M., Lammertyn, J., Verbruggen, F., and Vandierendonck, A. (2006). Tscope: a C library for programming cognitive experiments on the MS Windows platform. *Behav. Res. Methods* 38, 280.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., and Richmond-Welty, E. D.

(1995). The special role of rimes in the description, use, and acquisition of English orthography. *J. Exp. Psychol. Gen.* 124, 107–136.

van Hell, J., and de Groot, A. M. B. (1998). Disentangling context availability and concreteness in lexical decision and word translation. *Q. J. Exp. Psychol. A* 51, 41–63.

van Hell, J., and Dijkstra, T. (2002). Foreign language knowledge can influence native language performance: evidence from trilinguals. *Psychon. Bull. Rev.* 9, 780–789.

Yap, M. J., and Balota, D. A. (2009). Visual word recognition of multisyllabic words. *J. Mem. Lang.* 60, 502–529.

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: a new measure of orthographic similarity. *Psychon. Bull. Rev.* 15, 971–979.

Ziegler, J. C., Ferrand, L., Jacobs, A. M., Rey, A., and Grainger, J. (2000). Visual and phonological codes in letter and word recognition: evidence from incremental priming. *Q. J. Exp. Psychol. A* 53, 671–692.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 June 2010; paper pending published: 16 June 2010; accepted: 28 September 2010; published online: 18 November 2010.

Citation: Keuleers E, Diependaele K and Brysbaert M (2010) Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. Front. Psychology ***1****:174. doi: 10.3389/fpsyg.2010.00174*

This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.